A Distant Supervised Approach for Relation Extraction in Farsi Texts

Shireen Atarod
Faculty of Mechanics, Electrical Power and Computer Science and Research Branch Islamic Azad University
Tehran, Iran
At.shi.1989@gmail.com

Alireza Yari*
Assistant Professor
ICT Research Institute (ITRC)
Tehran, Iran
A_yari@itrc.ac.ir

Received: 2020/10/16 Revised: 2021/03/02 Accepted: 2021/04/02

***Abstract*— The volume of Farsi information on the Internet has been increasing in recent years. However, most of this information is in the form of unstructured or semi-structured free text. For quick and accurate access to the vast knowledge contained in these texts, the information extraction methods are essential to generate knowledge bases. In recent years, relation extraction as a sub-task of information extraction has received much attention. While many of these systems were developed in English and other well-known languages, the systems for information extraction in Farsi have received less attention from researchers. In this systematic research for semi-automatic relation extraction, Persian Wikipedia articles were presented as reliable and semi-structured sources. In this system, the relation extraction is performed with the assistance of patterns that are automatically obtained with an approach based on distant supervised. In order to apply the distant supervised, the vast knowledge base of Wikidata has been used as a source in perfect synchronization with Wikipedia. The results show that the average precision value for all relations is 76.81%, which indicates an enhancement of precision compared to other methods in Farsi.**

Keywords*— *Relation Extraction; Information Extraction; Distant Supervision; Persian Wikipedia

1. INTRODUCTION

Nowadays, the knowledge bases play an essential role in the structure of the semantic web. Therefore enriching these knowledge bases is important. Available sources include structured data such as Wikidata and Yago or semi-structured data such as Wikipedia InfoBox. However, many facts are found in unstructured texts. Relation extraction is used to extract these facts in the unstructured texts as the main component of information extraction.

In recent decades, the increasing development of Farsi textual information available on the Internet has been observed in various fields. For instance, the articles of Persian Wikipedia were improved remarkably in recent years both in terms of the number of articles and content richness. However, as mentioned previously, most of this information is in the form of unstructured texts, and the extraction of structured Farsi knowledge should be assisted by the relation extraction from these unstructured texts. Many studies have been performed on relation extraction in English and other languages in the world, but since relation extraction is heavily dependent on the syntax of the language of text, there is a need to perform research in other languages such as Farsi. In the field of natural language processing, Farsi is a less well-studied rich language, and few tools and resources are available for that. That is while a total of 1.5% of the world population are Farsi-speaking people [1].

Primary information extraction systems were particular to specific information and had been applying to a specific domain with similar texts and limited relations. Knowledge-based approaches work based on a type of pattern matching and rely on rich patterns to identify relations. These patterns can be a set of lexico-syntactic patterns [2] or manually designed for a specific event or subject [3, 4]. Although knowledge-based approaches can be quick to conclude, they are not easily transferable to other domains and require a great deal of manual work.

To pass through knowledge-based approaches to move toward trainable systems, the focus was on machine learning approaches, which are generally divided into three major categories, supervised, unsupervised, and semi-supervised. In supervised approaches [1, 5-7], a corpus is manually tagged as training data and classification is used to identify the type of extracted relations. Supervised methods depend on a set of tutorials that include tagged examples from a specific domain.

In unsupervised approaches [6, 8, 9], the text is considered as a sequence of words, and the types of relations are identified using statistical and clustering methods. These methods cannot detect relations between pairs of entities, and since there is no relation tag on each resulted instance from clustering, their results cannot be directly applied to knowledge bases.

For many language processing tasks, including relational extraction, there is a lot of unlabeled data, but labeled data is scarce and expensive to generate in large volumes, so it is desirable to develop automated techniques [10]. Semi-supervised methods use a limited amount of labeled data with a large amount of unlabeled data [11]. Since the labeling of samples in the field of natural language processing is very time-consuming, the use of such a method is very noteworthy [10]. As a result, using the small number of prototypes used to learn the pattern significantly reduces the required human effort. The main difference with supervised learning is that the act of tagging and preparing training data is done automatically rather than manually [12].

Semi-supervised methods can be divided into different categories depending on how the seeds are collected. These include bootstrap approaches [5, 13], open information extraction methods [14-16], self-supervised systems [1, 7, 17, 18], and distant supervised approach [3, 19-25].

The distant supervised approach is an efficient method for extracting relationships in large bodies that contain thousands of relationships [25]. The main idea of distant supervised is to use large databases to automatically label entities in the text, so that it can then use the annotated text to extract features and classify training [26]. The distant supervised method is used as automatic labeling of text with properties and resources, which are the sources of entities from a database [27]. The method of working in these methods is that if two entities are in a relationship, each sentence containing these two entities may express this relationship [27]. Since these methods do not require manual tagged datasets and knowledge bases have been growing recently, they have attracted a lot of attention [25]. In a recent research in Persian [28], they use FarsBase [24] and align entities in Persian Wikipedia articles to these relation instances and create a distantly supervised dataset.

This research presents a complete framework in Farsi for relation extraction from unstructured text. It is a distant supervised approach based on pattern matching. In this study, patterns extracted automatically at the entity level in Farsi. After automatic pattern extraction and storing them in a Pattern Base, they are used for extracting new relation instances from unstructured texts in Persian Wikipedia. The focus of the system is on seven relationships about individuals. These seven relationships are date of birth, gender, father, mother, sibling, spouse, and children.

After a brief description of the relation extraction, the necessity of the research, and a brief review of the different strategies proposed in other research in the first section, the paper continues as follows. In section 2, the previous works on semi-supervised approaches have been reviewed. The proposed method is outlined in section 3. Moreover, the implementation is described in section 4. Afterward, in section 5, the evaluation of the system has been reviewed, and finally, the conclusion with some suggestions for future works comes in section 6.

2. RELATED WORK

Mintz et al. [21] introduced the term distant supervised and implemented a system based on it for the first time. Distant supervised approaches are an efficient method for relation extraction in the large-scale corpus containing thousands of relations [29], which employ a vast knowledge base for applying supervision. In this algorithm, Freebase was used as the source of relations, and Wikipedia was employed as the subject of corpus [11]. For each pair of entities within Freebase relations, all sentences, including those entities, were identified from large untagged corpus [19]. The features syntactic, lexical, and Named-entity tag were used to generate patterns, and lexical features of the words between and around two entities were identified.

Gu et al. [30] introduced a new framework named Athena, which deals with relation extraction using a distant supervised approach at the entity level. Athena matches Freebase data with Wikipedia article to apply distant supervised for disambiguating the text by mapping the strings of words to entities. This research also uses the sequence of entities, the keywords between two entities, the N-gram sequence of essential keywords in the other two entities, and the NER and POS tags to generate patterns. It also uses a model based on Markov Logic Networks for relation extraction.

Although much research has been performed in English, they are hardly adaptive to other languages. To tackle this issue, in recent years, many other researchers tried to propose multi-lingual or non-English language approaches as shown in Table 1. Heist and Paulheim introduced a language-agnostic approach in [33]. This work implied a distant supervised approach using DBpedia knowledge-graph and Wikipedia abstracts as corpus. Instead of extracting information via language-dependent patterns, certain patterns discovered from the written structure of Wikipedia abstracts leveraging DBpedia background knowledge. Relation extraction demonstrated in the twelve languages of Wikipedia, using a Random Forest classifier.

In [32], Huang et al. introduced a multi-lingual approach, using a distant supervised model. This model works in both Chinese and English language. For distant supervised method, the Baidu encyclopedia is used and retrieved triplets from which matches with POS and NER tagged sentences by pattern matching based on regular expressions.

The works performed in Farsi is much less than the number of studies conducted in English and possess many limitations. Many of these studies define patterns manually and combine different strategies for information extraction.

One of the first works in the Persian language is the "Hasti" system [33]. Hasti uses both lexical-syntactic patterns and semantic patterns to derive semantic relations. In this research, models for extracting knowledge from Persian language have been introduced and their performance has been evaluated. Another system implemented in Persian is Mersad [3]. This system is designed to extract information from Persian news texts in the military field. For this purpose, the system uses information extraction patterns and provides a summary of the news to the user.

Sharifzadeh proposes a system for extracting information from Persian texts in a specific domain which can automatically extract the information contained in terrorist news [25]. The output of this system can be used to summarize news, extract conceptual knowledge of news, collect specific statistical information about events and many other applications.

Sudachi Khalese and Zare Bidaki [12] used a self-supervision method for general information extraction from the

TABLE 1. REVIEW OF DIFFERENT APPROACHES

Approach		Method	Pros	Cons
Knowledge-based		Domain independent	Fast results in the limited domain	Only specific domains A lot of manual efforts
		Domain-dependent		
Machine learning	Supervised	Feature based	Can be used in other domains	Act at the text level Heavy calculations Do not use tagged data
		Kernel based		
	Semi-supervised	Automatic start	Low volume training dataset General knowledge domain The least human effort	Error propagation Semantic change
		Extract free information		
		Self-supervision		
		Distant Supervised		
	Unsupervised	-	No need for tagged data	Not applied for pairs of entities

Persian Wikipedia texts to produce a knowledge base for the first time. The extractor consists of three components, preprocessor, matcher, and learner, and extracts the information by matching Wikipedia sentences with InfoBox attributes. Afterward, it fixes information defects by employing Freebase.

Another system [22] was also presented based on Wikipedia tacit knowledge, which only extracts semantic relations in Farsi documents. A weighted concept space is created for each phrase using this taxonomy feature in Wikipedia and semantic relation between two phrases as well as between two texts is calculated by employing the weighted concept space. The system in [34] is also based on the tacit knowledge of Wikipedia, which only extracts semantic relationships in Persian documents. Wikipedia has a hierarchical structure called a category that organizes all articles into a hierarchy [33]. Using this feature, it creates a weighted concept space for each phrase, and using the weighted concept space, the semantic relationship between the two phrases as well as between the two texts is calculated. Fapedia [35] is a Persian cognitive database extracted from Wikipedia. Cognition database extraction is a special type of information extraction and its purpose is to extract a hierarchical construction of concepts and the most common relationships between them [35]. WikiInfo system [36] is a self-monitoring system for producing and completing information boxes of celebrities' pages in Persian Wikipedia.

Another system introduced for relation extraction in Farsi contents is a system in which extracts the information of persons in Farsi contents by employing patterns. The relevant information of the focus persons is identified and extracted in the corpus. For this purpose, pattern matching consisting of a set of keyword-based patterns, and the value for predefined candidate relevant information has been used [1]. This research was conducted on ten relations and has obtained a list of values for candidate relevant information manually from sources such as Wikipedia.

So far, not much research has focused on relationship extraction in Persian texts. In Persian language re-searches, one or more steps of the work are done manually, and, in many cases, combined approaches are used to extract relationships. In the latest works in Farsi, Nasser et al. [23] introduce a distant supervised approach for relation extraction on Persian Wikipedia articles. For applying supervision, the FarsBase [24] is used. For extracting new relation instances, a Piecewise Convolutional Neural Networks (PCNN) model is used, and features are extracted automatically from a sentence that a retrieved pair of entities from knowledge graph appears in it. over 8000 relations [23]. RePersian [28] is dependent on part-of-speech (POS) tags of a sentence and special relation patterns, which are extracted by analyzing sentence structures in Persian. Perlex [37] is introduced as the first Persian dataset for relation extraction, which is an expert translated version of the “Semeval-2010-Task-8” dataset. FarsBase is a Persian knowledge graph which constructed from various sources and contains over 7.5 million relation instances.

In this paper, a distant supervised approach like [28] is introduced for automatic relation extraction on the unstructured text at the entity level in Farsi. This performed on Persian Wikipedia articles and Wikidata applied as supervision. Relation tuples from Wikidata matched with Wikipedia sentences and automatically extracted relation patterns stored in a Pattern Base. Then new relation tuples extracted from unstructured text via these patterns automatically. To train the system, data samples from the vast database of Wikidata have been used. First, prototypes of the Wikidata server for each relationship are collected. Then, for each relationship, the texts of Persian Wikipedia articles were matched with the collected samples, and the patterns of each relationship were automatically extracted. Then, by matching these patterns (using regular noun phrases) with the texts of Wikipedia articles, new samples are automatically extracted.

3. THE PROPOSED SYSTEM

In [21] distant supervised paradigm described as follows, "If two entities participate in a relation, any sentence that contains those two entities might express that relation." Hence, for each instance of a relation, all the text searched for those two entities and the found sentence is considered as a candidate as a specific pattern for that relation. After a few tagging and processing steps, the sentence will be transformed into a general pattern and if it has not been found yet, it will be saved in the PatternBase. After all, for finding new instances for each relation, all sentences of an unstructured text will be matched with all patterns of that relation in PatternBase. And in case of matching, the entities in the matched sentence shall be added to the knowledge base.

In this research, an attempt has been made to implement an automatic extraction of relationships from Persian Wikipedia articles using an algorithm based on the distant supervised approach. The focus of the system is on seven relationships about individuals. These seven relationships are date of birth, gender, father, mother, sibling, spouse, and children. The system uses a maximum of 1000 prototypes for each relationship and 50% is considered as training data and the other 50% as test data. For each relationship, the text of the articles corresponding to the first existence of the training data samples were isolated and the automatic pattern extraction operation was performed on them. The system then extracts new samples from the articles corresponding to the test samples and compares the result with the test dataset for evaluation. Although the scope of this system is limited to a specific area, it can perform relationship extraction in general.

The proposed system consists of three main modules, preprocessing module, pattern extraction module, and relation extraction module (Fig. 1). The system possesses three input data, Wikipedia as a corpus, Wikidata as a knowledge base, and the set of relations consists seven relations. Wikipedia characteristics no longer raise problems such as homonyms and synonym since every concept in Wikipedia is mapped to a unique URI. On the other hand, Wikidata is in perfect synchronization with Wikipedia as one of the Wikimedia projects, and each entity is indented in all versions of Wikipedia in different languages with the same URI. These entities also have some statements, which are also specified by a URI. Each of these statements represents one type of relations. Hence, each entity E1 (primary entity) be in relation r with the E2 (second entity), is considered as a r (E1, E2) tuple of relation of r.

In the first step, the preprocessor module was provided with the dump of Persian Wikipedia articles which is not prepared for processing. After a few steps of preprocessing and data

cleaning the plain text would be ready. Then patterns were extracted automatically at the entity level in the pattern extraction module from plain texts by mapping the relation instances obtained from Wikidata. In pattern-based methods, input (usually text) is searched for a specific pattern or keyword that represents a particular conceptual relation. In this system, patterns are extracted at a higher level than the textual level by matching the string of words in the text with the entities within the knowledge base and disambiguating them. The extracted patterns are stored in a pattern base, and finally, the plain text's sentences were matched with all the patterns. New tuples for specific relations were extracted using their corresponding patterns, and the results were stored in a knowledge base. The output of the system of pattern bases and the set of new tuples were extracted.

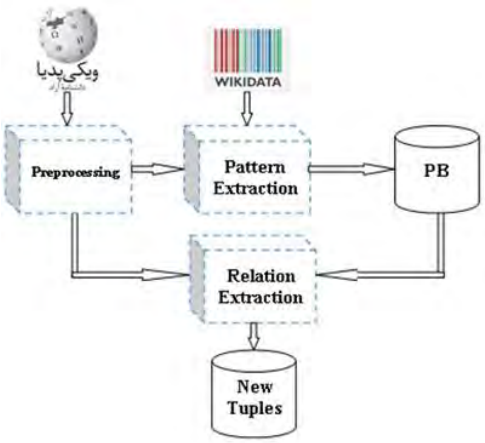

Fig. 1. System framework

3-1 Preprocessing Module

As shown in Fig. 2, the preprocessing module has three components. The input to the first component is the dump of Wikipedia articles which is in XML format and not prepared for text mining; therefore, in this stage of preparation, non-article pages that are part of the Wikipedia structure should be removed (such as talk pages, disambiguate pages, redirect pages, wiki documentation, etc.). Afterward from the remaining articles, all the XML, HTML and Wikipedia specific tags are eliminated from the text except one special tag. In Wikipedia articles, entities surround in a "[[]]" which aims to process text at entity level since there is no effective tool for entity recognition in Farsi. Finally, the prepared plain text of the articles remains.

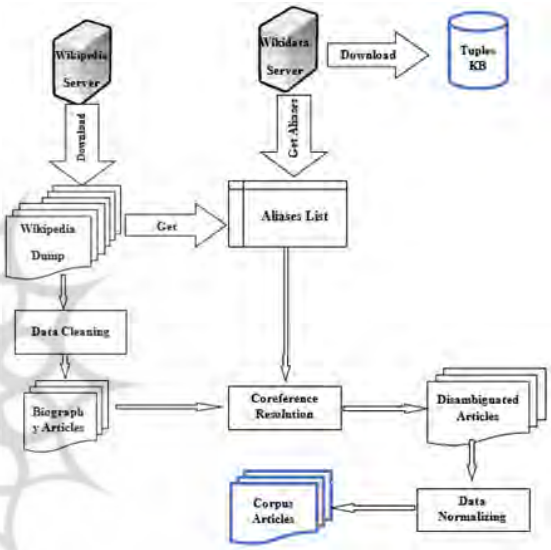

Fig. 2. Preprocessing module

The next step is to disambiguate entities in the texts. Ambiguity in the text occurs when different textual forms, aliases, or even pronouns that refer to the main entity in sentences, cause problems in correctly identifying the entity. Since there is no efficient co-reference resolution tool for Farsi Language available, dis-ambiguation to the original entity of the article is performed in this study using a two stages heuristic approach.

In the first stage, the textual disambiguation and aliases were discussed. Textual ambiguities occur when the name of an entity is spelled differently. For example, "تهران" and "طهران" both mean "Tehran". On the other hand, an alias like "تهران بزرگ" (Great Tehran) may be used instead of "تهران" (Tehran) in the text. Since each topic of Wikipedia article possesses a unique URI on Wikidata, a list of aliases and different textual forms has been provided for each URI. Two different sources were used to provide the mentioned list. The first source is the redirect pages, which are of structural features of Wikipedia. In some previous research, such as [25], the redirect pages were used for co-reference resolution. The second source is the aliases list on Wikidata, which contains a list of different names and different textual forms of each entity. In the present study, a list of co-references for each entity was prepared by integrating these two sources. All these aliases were replaced with the original entity name in the preprocessing stage.

In the second step, the disambiguation of pronouns in the text was discussed. The pronouns in Farsi language can be in the connected form or disconnected form. Connected pronouns are all used for conjugation and do not cause any ambiguity in the sentence. However, disconnected pronouns can be in the role of subject or object and appear anywhere in the sentence.In the present study, it was assumed that all subjective or objective pronouns in the text of an article refer to the entity of the owner of that page. Hence, these pro-nouns were also changed to the original entity name. This assumption can cause some errors in the text, but its general precision is acceptable.

The final step of preprocessing is the text normalization. One of the essential practices at this stage is unification of the word's boundary. Space is one of the most important factors of ambiguity in Farsi language because the letters of Farsi alphabet forming a single word are written as clinging form; however, if a word is a combination of several components, these components can be written with space or semi-space. For instance, both the terms "کتاب‌ها" and "کتاب ها" mean "books", in which semi-space and space were used in the first and second phrases, respectively. In Farsi writing, space and semi-space possess two different Unicode that can interfere with automatic text processing. For this reason, all spaces were converted to semi-space since semi-spaces make processing easier. Diacritics are of other characters that cause ambiguity in the text (fatha, k asra, damma). Diacritics in the Farsi texts only contribute to the fluency of the text, but they cause two identical strings to be separately recognized when compared.

For this reason, all diacritics were removed from the texts. Finally, Unicode unification of Farsi and English numbers is done, and all English numbers are converted to Unicode of Farsi Numbers. Ultimately, after completing these steps, the articles are ready to be processed in plain texts.

3-2 Pattern Extraction Module

In the pattern extraction module, pattern extraction in the text was discussed so that new tuples could be extracted using the patterns as shown in Fig.3.This module runs once per relation. According to the definitions mentioned in [19], patterns are the sequence of words of essential keywords. "Essential keywords" are words that their presence in the considered sentence is somehow necessary for the relation. In this system, TF-IDF has been used for term weighting to generate N-gram patterns.

For the pair of entities of assumed relation, $X_i = r(E_1, E_2)$, the Wikipedia page of primary entity E_1 is extracted from the preprocessed corpus. Then, each sentence of the article is matched with X_i tuple; it means that if both Xi entities exist within a sentence, the sentence will be considered as a candidate for pattern extraction. Then, the words that form the candidate sentence are tagged by the named-entity recognition and the part-of-speech tagging systems. For named-entity tagging a NER tagger is used that designed for Farsi and tag entities as "per" (person), "loc" (location), "org" (organization) and "o" (others). Also, for tagging "Date" entities, a heuristic method used. In this method, each entity that contains numerical characters plus one of the dating system indications consider as a Date entity.

The obtained patterns include a flag to indicate the order of the entities in the sentence, the named-entity tag of entities, the sequence of words between two entities, the sequence of words in the window of k words to right side of first entity (since Farsi is a right-to-left language), the sequence of words in the window of k words to left side of second entity, and the verb at the end of sentence. Since the verb always comes at the end of Farsi sentences, the verb must be independently a part of the produced pattern.

For instance, the phrase "[[E1]] ([[E1]] مادرِ [[E2]] است." ([[E1]] is [[E2]]'s mother.) is an example of a possible pat-tern for "mother" relation. It means that the pattern "per-[[…]] (Verb) است per-[[…]] مادرِ" ([[…]]-per is [[…]]-per's mother.) is essential as a sequence of words to reach the "mother" relation (as a type of relation).

3-3 Relation Extraction Module

The components of this module are shown in Fig. 4. The module runs once for each relation. First, the text of each entity article is divided into sentences. Afterward, if a sentence contains the name of the article's entity E1, it is assumed that the sentence may contain that relation. The components of the sentence are tagged with the NER and POS taggers. Afterward, the tagged sentence matched with all the patterns of interested relation. If the sequence of entities, sequence of words between and outsides of entities, and the named-entity tag of the entities completely match the pattern, consider the second entity E2 as the secondary entity and stored the tuple (E1, E2) as a new tuple for the relation in the knowledge base.

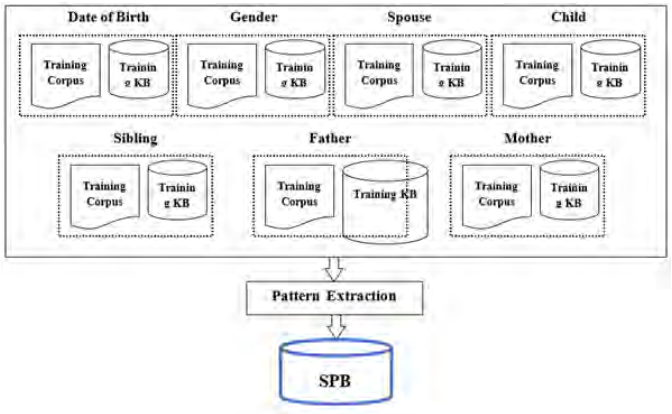

Fig. 3. Pattern extraction module

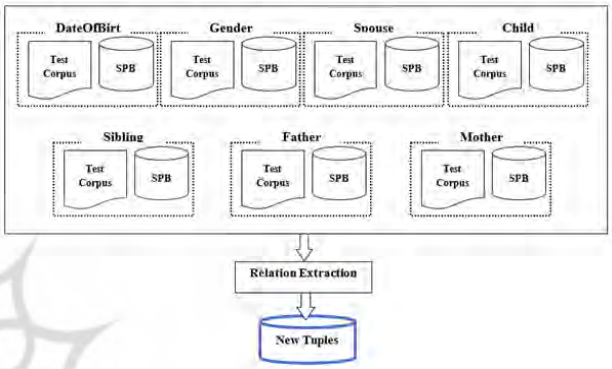

Fig. 4. Relation extraction module

4. IMPLEMENTATION

4-1.Data

The full version of Persian Wikipedia to date 2017/06/20 in XML format has been used for the present study, which includes 2,609,485 pages before the preprocessing stage. To create a knowledge base for each relation, the maximum of 1000 seeds were considered. The system test was carried out on seven relations, including "date of birth, gender, father, mother, siblings, spouse, and child". This system was performed on Ubuntu 14.06.3, and the Intel® Core ™ i7-4600 processor and 2GB of RAM were installed on the VMware virtual machine.

Table 2 represents the number of patterns extracted for relations. Among the seven relations mentioned above, the "gender" relation follows a different pattern. Since there is no clear indication for expressing the gender of subject in the sentence in Farsi grammar, two lists were prepared for male and female names, which included 1377 male and 821 female names. 234 male and 41 female names were added to the lists during pattern extraction. In relation extraction module, entity's name is searched in each list and if matches, a new tuple of entity and its gender is added to the knowledge base.

5. EVALUATION

Precision and recall criteria have been used in most research in the field of information extraction and relation extraction to evaluate the proposed systems. Precision is the percentage of related instances among the retrieved instances,

and recall is the percentage of related instances that have been retrieved. Table 3 represents the precision and recall of each relation. In summary, the precision value for the seven considered relations is in the range of [57% - 98%].

Moreover, the recall value is in the range of [2.8% - 23.6%]. Many researchers find it challenging to compute recall (and sometimes even precision) in all semi-supervision algorithms. Since it is difficult to obtain the precise number of entities of the relation in large volumes of data, it is also difficult to calculate recall to evaluate supervision methods [20]. In other words, the number of instances that existed for a relation is not clear.

There has been much research in the field of relation extraction with self-supervision and distant supervised approaches in the English language. However, such studies are very limited in Farsi language, and in many cases, quantitative criteria such as precision and recall have not been presented to evaluate the sys-tem.

For comparison, the results have been compared with the other methods in the Table 4. The Fadaei system [38] has performed information extraction from the Persian Wikipedia using a supervised approach and a set of manually crafted patterns. The Emami system [1] has also employed a semi-supervised approach in Farsi corpus and a set of patterns collected manually.

The Nasser system [23] has performed a distant supervised approach for relation extraction from Persian Wikipedia articles using a neural network with automatically extracted eatures. Following [23], RePersian [28] is presented by using FarsBase [24] and align and align entities in Persian Wikipedia articles to these relation instances and create a distantly supervised dataset. Athena [30] has performed automatic pattern extraction and relation extraction from English Wikipedia by a distant supervised approach. Fig.2 indicates the average precision obtained in each system. The comparison of the average precision of different systems is depicted in Fig.5.

In most research in Farsi language, all, or part of the process of collecting data or pattern sets has been performed manually, which requires a great deal of time and effort. Generally, evaluations indicate that the proposed system significantly reduces the human effort for data collection and pattern extraction, in addition to improving precision and recall percentages compared to previous works in Farsi language. Also, its range of operation is not limited to a specific domain, and if there are an appropriate corpus and knowledge base, it is responsive for extracting instances of any relation. No matter how promising the semi-supervised methods are, error propagation is a severe problem of these methods, an error in earlier steps causes more problems in the next steps and reduces the accuracy of the extraction process. Other problems with these methods include the problem of semantic change, the assumption of the existence of only one relation between two entities, etc. There are various problems in Farsi language as well, such as lack of data sources, incomplete Farsi data on sources such as Wikidata, ambiguity in Farsi language, lack of efficient natural language processing tools for Farsi language, the inaccessibility of published research papers and documents, were of the main obstacles to the present research.

TABLE 2. NUMBER OF EXTRACTED PATTERNS

Relation	**#Patterns For All K**
DateOfBirth	**387**
Spouse	**192**
Child	**51**
Sibling	**87**
Father	**156**
Mother	**18**
Gender (Male/ Female)	**275**
	234

TABLE 3. PERFORMANCE FOR ALL RELATION TYPES

Relation	**Precision**	**Recall**
Date of Birth	63%	**23.6%**
Gender	98%	**39%**
Spouse	85.7%	**2.8%**
Child	81%	**10.1%**
Sibling	75%	**4.8%**
Father	78%	**6%**
Mother	57%	**4.5%**

TABLE 4. PERFORMANCE FOR ALL APPROACHES

Research	Approach	Data	Language	Collection Method	Average Precision
Fadaei	Supervised	Wikipedia	Farsi	Manual	44%
Emami	Semi-supervised	Wikipedia Tabnak	Farsi	Manual	21%
Athena	Distant Supervised	Wikipedia	English	Automatic	84.1%
RePersian	Distant supervised	FarsBase	Farsi	Automatic	78.05%
Nasser	Distant supervised	Farsbase	Farsi	Automatic	76.81%
Proposed system	Distant supervised	Wikipedia	Farsi	Automatic	76.81%

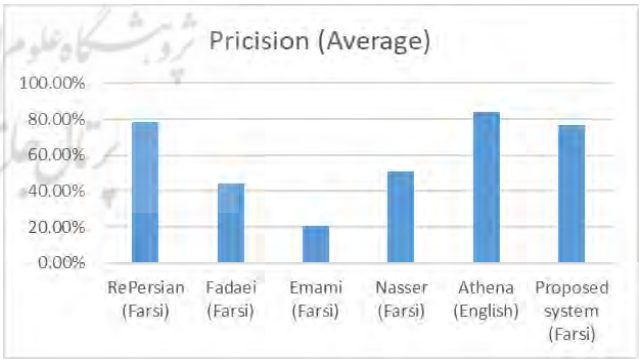

Fig. 5. comparison of average precision of different systems

6. CONCLUSION

In this research, an attempt has been made to implement a system for automatic relation extraction based on a pattern from Persian Wikipedia articles using an algorithm based on the distant supervised approach. Wikipedia has been used as a corpus and Wikidata as a knowledge base in full synchronization with Wikipedia to apply supervision for the training of the extractor, using the seeds in the vast knowledge

base of Wikipedia. First, the primary instances were collected from the Wikidata server for each relation. The texts of Persian Wikipedia articles are then matched with the collected instances for each relation, and the patterns of each relation were automatically extracted. Then, new instances were automatically extracted by matching these patterns with the texts of Wikipedia articles. This system can extract general relations. The results also demonstrate that the precision obtained in this study was significantly improved for the extracted in-stances by 76.81% compared to other studies in Farsi.

Problems such as lack of data sources, incomplete Persian data in resources such as Wikidata, ambiguity in Persian language, lack of tools for efficient natural language processing for the Persian language, and of course, the unavailability of published documents and research works are the main obstacles.

As mentioned above, the lack of Wikipedia articles and Wikidata samples makes the training difficult. Therefore, using several other data sources and combining their information with Wikipedia and Wikidata can significantly help to improve the training process.

Another limitation in this research was the lack of appropriate preprocessing tools for the Persian language. The lack of an efficient system for determining the nominal phrases of the reference at the time of implementation. Therefore, future researchers are advised to use an efficient module for this purpose. Also, numerous errors were observed in the tools of labeling entities, and with the development of these tools, one can expect an improvement in the values of precision and recall in future works.

REFERENCES

[1] H. Emami, H. Shirazi, A. Abdollahzadeh, and M. Hourali, "A Pattern-Matching Method for Extracting Personal Information in Farsi Content", *University Politehnica of Bucharest-Scientific Bulletin, Series C, Electrical Engineering and Computer Science*, vol. 78, pp. 125-139, 2016.

[2] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora", In *Proceedings of the 14th conference on Computational Linguistics*,Vol. 2, 1992, 539-545.

[3] N. Rahimipour, M. Shamsfard, and Z. Ansari, "Information Extraction System, Mersad", In *Proceedings of the fifteenth Iran conference on Electric Engineering*, 2007.

[4] A. Sharifzadeh and M. Shamsfard, "Automatic Information Extraction on Special Domain", In *Proceedings of nineteenth Annual National Conference on Iran Computer Society*, 2014.

[5] S. Brin, "Extracting patterns and relations from the World Wide Web", *International Workshop on The World Wide Web and DatabasesSpringer*, Berlin, Heidelberg ,1998, pp. 172-183.

[6] J. Chen, D. Ji, C. L. Tan, and Z. Y. Niu, "Relation extraction using label propagation based semi-supervised learning", In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, 2006, pp. 129-136.

[7] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Web-scale information extraction in knowitall, (preliminary results)", In *Proceedings of the 13th international conference on World Wide Web*, ACM, 2004, pp. 100-110.

[8] R. Feldman and B. Rosenfeld, "Boosting unsupervised relation extraction by using NER", *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006, pp. 473-481.

[9] T. Hasegawa, S. Sekine, and R. Grishman, "Discovering relations among named entities from large corpora", In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*Association for Computational Linguistics, 2004, pp. 415-422.

[10] N. Bach and S. Badaskar, "A review of relation extraction", *Literature review for Language and Statistics II*, vol. 2, pp. 1-15, 2007.

[11] R. Grishman, *Information Extraction: Capabilities and Challenges*, Lecture Notes of Computer Science, 2012.

[12] P. Sudachi Khalese and M. A. Zare Bidaki, "An information framework for automatic answering to Farsi questions based on extracted knowledge from Wikipedia using self-supervised learning", In *Proceedings of 3th International Conference on Applied research in Computer and Information*, 2016.

[13] E. Agichtein and L. Gravano, "Snowball, Extracting relations from large plain-text collections", In *Proceedings of the fifth ACM conference on Digital libraries*, ACM, 2000, pp. 85-94.

[14] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the web", *IJCAI*. Vol. 7, pp. 2670-2676, 2007.

[15] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction", *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 1535-1545.

[16] F. Wu and D. S.Weld, "Open information extraction using Wikipedia", In *Proceedings of the 48th annual meeting of the association for computational linguistics*, Association for Computational Linguistics, 2010, pp. 118-127.

[17] B. Rozenfeld and R. Feldman, "Self-supervised relation extraction from the Web", *Knowledge and Information Systems*, vol. 17, no. 1, pp. 17-33, 2008.

[18] D. S. Weld, F. Wu, E. Adar, S. Amershi, J. Fogarty, R. Hoffmann, and M. Skinner, "Intelligence in Wikipedia", In *AAAI*, vol. 8, pp.1609-1614, 2008.

[19] N. Konstantinova, "Review of relation extraction methods, what is new out there?", *International Conference on Analysis of Images, Social Networks and Texts*, Springer, Cham, 2014, pp. 15-28.

[20] B. Min, R. Grishman, L. Wan, C. Wang, and D. Gondek, "Distant supervision for relation extraction with an incomplete knowledge base", In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies*, 2013, pp. 777-782.

[21] M. Mintz, S .Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data", In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, vol. 2, Association for Computational Linguistics, 2009, pp. 1003-1011.

[22] A. Mosalla Nejad, D. Davoodi Moghadam, and A. Ahmadi, "An effective algorithm for semantic relation extraction in documents based on Wikipedia knowledge base", *Proceedings of 23th Iran Electrical Engineering Conference*, 2016, pp. 918-923.

[23] M. Nasser, M. Asgari, and B. Minaei-Bidgoli, "Distant Supervision for Relation Extraction in The Persian Language using Piecewise Convolutional Neural Networks", *5th International Conference on Web Research (ICWR)*, IEEE, 2019, pp. 96-99.

[24] G. Ji, K. Liu, S. He, J. Zhao, "Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions", In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017.

[25] K. Xu, S. Reddy, Y. Feng, S. Huang, and D. Zhao, "Question answering on freebase via relation extraction and textual evidence", arXiv preprint arXiv:1603.00957, 2016.

[26] S. Heydari, Z. Banaian, and V. Reshadat, "Study of information extraction methods based on machine learning and knowledge engineering", *The Second International Conference on Knowledge-Based Research.* Tehran, Majlisi University, 2017.

[27] R. Saheb-Nassagh, M. Asgari, and B. Minaei-Bidgoli, "RePersian A Fast Relation Extraction Tool in Persian", *International Journal of Web Research*, vol. 2, no. 2, Autumn-Winter, 2019.

[28] M. Asgari-Bidhendi, A. Hadian, and B. Minaei-Bidgoli, "FarsBase: The Persian Knowledge Graph", *Semantic Web*, voll. 10, no 6, IOS Press, 2019.

[29] Y. Gu, W. Liu, and J. Song, "Relation extraction from Wikipedia leveraging intrinsic patterns", *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, IEEE, 2015, pp. 181-186.

[30] N. Heist, S. Hertling, and H. Paulheim, "Language-agnostic relation extraction from abstracts in Wikis", *Information*, vol. 9, no. 4, p. 75, 2018.

[31] Y. Huang, Y. Jia, J. Huang, and Z. He, "Multi-language person social relation extraction model based on distant supervision", *IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*. IEEE, 2018, pp. 386-374.

[32] M. Shamsfard and A. Abdollahzadeh Barforosh, "Extracting conceptual knowledge from the text using linguistic and semantic patterns", *Cognitive Science News*, vol. 4, no 1, pp. 48-60, 2002

[33] A. Mosallanejad, J. Davoodi Moghadam, and A. Ahmadi, "Presenting an efficient algorithm for extracting semantic relationships in documents, based on the tacit knowledge base of Wikipedia". *23rd Iranian Electrical Engineering Conference. Tehran, Sharif University of Technology*, 2015.

[34] S. Dami, H. Shirazi, and A. Abdullah Zadeh, "Fapedia, a large-scale Persian cognitive database extracted from DBPedia", *4th Joint Congress of Fuzzy and Intelligent Systems of Iran*, Zahedan, University of Sistan and Baluchestan, 2015.

[35] M. Asgari-Bidhendi, M. Nasser, B. Janfada, and B. Minaei-Bidgoli, "Perlex: A Bilingual Persian-English Gold Dataset for Relation Extraction", *Scientific Programming*, 2020.

[36] P. Khaleseh Sudachi, "Automatic production of Persian information boxes for individuals using the extraction of information made from Wikipedia articles", *The first national conference on new ideas in electrical and computer engineering*, Iran, 2016.

[37] D. Hasili, M. Hosseini Beheshti, and S. Pak Nohad, "Information Extraction, Methods and Applications", *The first international conference on interactive information retrieval*, Tehran, University of Tehran, 2016.

[38] H. Fadaei and M. Shamsfard, "Extracting conceptual relations from Persian resources", In *Proceeding of Seventh International Conference on Information Technology, New Generations*, 2010, pp. 244-248.

[39] N. Kambhatla, "Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations", In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 2004, pp. 178-181.

[40] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning, "Multi-instance multi-label learning for relation extraction", In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, Association for Computational Linguistics, 2012, pp. 455-465.

[41] D. Zelenko, C. Aone, and A. Richardella, "Kernel methods for relation extraction", *Journal of machine learning research*, vol. 3(Feb), pp. 1083-1106, 2003.

[42] S. Zhao and R. Grishman, "Extracting relations with integrated information using kernel methods", In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005, pp. 419-426.

Shireen Atarod received her bachelor's degree in information technology (IT) engineering from Hamedan University of Technology (HUT), From Hamedan, Iran, in 2012. She received her master's degree in e-commerce from Science and Research Branch of Islamic Azad University (SRBIAU) in 2018. Her research interests include relation extraction, supervised and semi supervised machine learning, and text mining.

Alireza Yari received his B.Sc. degree in control system engineering in 1993 from the University of Tehran, Iran, and M.Sc. and a Ph.D. degree in System engineering in 2000 from Kitami institute of technology, Japan. He is currently doing research in the Information Technology research faculty of Iran Telecom Research Center (ITRC). His research interests include web processing and cyber linguistics application, such as web search engines.